# Chapter_6

*Christophe*

*October 13, 2015*

## 6.6 2010 Healthcare Law.

**On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.**

**(a) We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.**

False, we are 100% confident that the Americans in the sample fall between 43% and 49%.

**(b) We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.**

True, the 3% margin of error allows us to construct a confidence interval of 43% to 49%.

**(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.**

True, the confidence interval implies that if many samples were made of the population 95% of the time the mean would fall between 43% and 49%.

**(d) The margin of error at a 90% confidence level would be higher than 3%.**

False, the margin of error is lower at a lower confidence interval because we become less confident that our confidence interval contains the true population mean but we are able to generate a smaller confidence interval.

## 6.12 Legalization of marijuana, Part I.

**The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.**

**(a) Is 48% a sample statistic or a population parameter? Explain.**

48% is a sample statistic because it is not an inference of the entire population but the statistic generated from the surveyed people.

**(b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.**

The confidence interval assuming a 95% confidence and a somewhat normal distribution would be ( 45.2% , 50.8% )

**(c) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.**

Not necessarily, because we have sampled well above the 30 samples threshold, skew is acceptable. We should always be concerned with strong skew though.

**(d) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified**

Sure, my confidence interval could be used either way since it straddles 50%. Since I support legalization I would find nothing wrong with this article. I would probaby have more issues if it was used to support the opposition.

## 6.20 Legalize Marijuana, Part II.

**As discussed in Exercise ??, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?**

$1.96 * sqrt(p * (1 - p))/n) < .02$

$1.96^2 * (p * (1 - p))/n < .02^2$

$1.96^2 * (p * (1 - p))/.02^2 < n$

$p = .48$

$2397.1584 < n$

$2398 < n$

So 2,398 people would need to be surveyed.

## 6.28 Sleep deprivation, CA vs. OR, Part I.

**According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.**

```
x <- sqrt(((.08 * (1-.08))/11545) + ((.088 * (1-.088))/4691))
y <- .08 - .088
high <- y + (1.96 * x )
low <- y - (1.96 * x )
paste("The confidence interval is", "(", percent(low), ",",percent(high),")")
```

[1] "The confidence interval is ( -1.75% , 0.15% )"

The confidence interval gives us sufficient information to see that difference in insufficient sleep is probably not due to location, and Califronians and Oregonians suffer from sleep issues at the same proportion.

## 6.44 Barking deer.

**Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests makes up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data**

| Woods | Cultivated grassplot | Deciduous forests | Other | Total |
|---|---|---|---|---|
| 4 | 16 | 61 | 345 | 426 |

**(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.**

$H_o$ : barking deer do not have a preference for certain habitats for foraging.

$H_a$ : barking deer do have a prefernce for certain habitats for foraging.

**(b) What type of test can we use to answer this research question?**

The type of test we can use to answer this question is the chi-squared goodness of fit test.

**(c) Check if the assumptions and conditions required for this test are satisfied.**

| | Woods | Cultivated grassplot | Deciduous forests | Other | Total |
|---|---|---|---|---|---|
| Expected | 21 | 63 | 169 | 175 | 426 |

The expected results of each scenario is greater than 5 and each case appears to be independent of the others. However, a biologist may have more insight if each case is truly independent. For the the sake of this exercises we will assume it is independent and the test for inference is satisified.

**(d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.**

```
woods_observed <- 4
woods_expected <- round((426 * .048) + .5, 0)
cultivated_gl_observed <- 16
cultivated_gl_expected <- round((426 * .147) +.5, 0)
deciduous_f_observed <- 61
deciduous_f_expected <- round((426 * .396) +.5, 0)
other_observed <- 345
other_expected <- round((426 * .409) +.5, 0)
Z_woods <- (woods_observed - woods_expected) / sqrt(woods_expected)
Z_cultivated_gl <- (cultivated_gl_observed - cultivated_gl_expected) / sqrt(cultivated_gl_expected)
Z_deciduous <- (deciduous_f_observed - deciduous_f_expected) / sqrt(deciduous_f_expected)
Z_other <- (other_observed - other_expected) / sqrt(other_expected)
Z_two <- Z_woods^2 + Z_cultivated_gl^2 + Z_deciduous^2 + Z_other^2
```

The $\chi^2 = 282.99$ and based on the degrees of freedom being `n - 1` we get `3` degrees of freedom. Using the chi-distribution we see the value of $\chi^2$ is greater than the value at `p = 0.001` since our $\chi^2$ is so much larger we reject the null hypothesis and conclude that barking deer do have a prefernce for certain habitats for foraging.

## 6.48 Coffee and Depression.

**Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption**

| | | *Caffeinated coffee consumption* | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\leq 1$ cup/week | 2-6 cups/week | 1 cup/day | 2-3 cups/day | $\geq 4$ cups/day | Total |
| *Clinical depression* | Yes | 670 | 373 | 905 | 564 | 95 | 2,607 |
| | No | 11,545 | 6,244 | 16,329 | 11,726 | 2,288 | 48,132 |
| | Total | 12,215 | 6,617 | 17,234 | 12,290 | 2,383 | 50,739 |

**(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?**

The chi-test test for two-way table.

**(b) Write the hypotheses for the test you identified in part (a).**

$H_o$ = There is no difference in clinical depression among women by their level of caffeine consumption

$H_a$ = There is some difference in clinical depression among women by their level of caffeine consumption

**(c) Calculate the overall proportion of women who do and do not suffer from depression.**

The proportion of women who suffer from depression is 5.14% and the proportion of women who do not suffer from depression is 94.9%.

**(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e.** $(Observed - Expected)^2/Expected$**.**

Assuming that caffeinated coffee consumption has no impact on clinical depression and the current proportion is normal than we multiply that proportion by the number of users in that category. Therefore the expected value is 337.467 or 338 people.

The test statistic for this cell is 3.62.

**(e) The test statistic is** $\chi^2 = 20.93$**. What is the p-value?**

We first need to determine the degrees of freedom which is $df = (row - 1)\ddot{O}(column - 1)$ so it would be `df = (2 - 1) x (5 - 1)` or df = 4.

Based on the degrees of freedom the p-value of the chi-test is greater than 14.860 which is the p value at .005.

**(f) What is the conclusion of the hypothesis test?**

The conclusion is that we reject the null hypothesis and that there is a difference of clinical depression among women according to their coffee consumption.

**(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study. Do you agree with this statement? Explain your reasoning.**

I would agree with this statement, this was simply an observational study so we can only make correlated inferences not causality inferences. There was not a control group accounting for the other confounding factors weo can while we can say that coffee consumption is correlated with clinical depression we cannot say that it has a causal affect on depression.